ニューラルネットワークに基づく発話動画生成に関する研究

修士学位論文要約（平成30年３月）

ニューラルネットワークに基づく発話動画生成に関する研究

佐藤 一樹
指導教員：伊藤 彰則，　研究指導教員：能勢 隆

A Study on Talking Head Synthesis Based on Neural Networks

Kazuki SATO
Supervisor: Akinori ITO,　Research Advisor: Takashi NOSE

“Talking Head Synthesis” is a technique to generate images of the face of a photo-real person who is speaking from inputted voice or text. In this paper, I consider the application to spoken dialogue systems, and propose the techniques to generate photorealistic talking head from texts. In previous researches about talking head synthesis, modeling on the dynamic features and parameter generation technique considering variance of parameters in one utterance have not been sufficiently studied, so in this paper, I introduce them to talking head synthesis and verify its effect. From results of experiments, I showed that the proposal technique using minimum generation error training and generative adversarial networks can generate more natural talking heads than the conventional techniques.

1.　はじめに

音声やテキストを入力とし，それを喋っているリアルな人の顔の映像を生成する技術は Talking head synthesis や Visual text-to-speech と呼ばれる．本稿ではこれらをまとめて発話動画生成と呼ぶこととする．発話動画生成は，音声合成技術と組み合わせることで任意の発話映像を生成することが可能となり，主な応用先として，音声対話システムや外国語教育システム，映像コンテンツ制作などが考えられる．

本研究では，音声対話システムへの応用を想定し，テキストから高精度でフィトリアリスティックな発話動画を生成する手法について検討する．特に本研究では，顔モデルとして Active Appearance Model (AAM) ，生成モデルに Deep Neural Network (DNN) を用いた手法に着目する．先行研究において十分に検討が行われていない，パラメータの動的特徴のモデル化や，一発話内におけるパラメータの分散を考慮したパラメータ生成手法に関して検討を行い，その効果を検証する．

2.　DNN に基づく発話動画生成

本研究では顔モデルに AAM，生成モデルに DNN を用いた発話動画生成について検討する．AAM は顔の形状と色それぞれに対し平均からの差分をパラメータ化することで未知の画像を表現する 2 次元顔モデルであり，一般的な RGB 動画像データからモデルの構築及びパラメータの算出が可能であるためデータセットの収録に特殊な機材が不要であるというメリットがある．

DNN に基づく発話動画生成では，入力文に対応するコンテキスト系列から画像特徴量系列へのマッピングを DNN によりフレーム単位でモデル化する．DNN の入力は先行，当該，後続の音素の種類と音素内相対位置を表すコンテキストベクトル，出力は AAM パラメータとその 1 次・2 次の動的特徴量である．動的特徴量は前後のフレームからの変動を表すパラメータである．DNN から出力された静的・動的特徴量系列から Maximum Likelihood Parameter Generation (MLPG) によって動的特徴を考慮した AAM パラメータを算出し，フレームごとに顔画像を再構成することで最終的な発話動画を得る．

3.　動的特徴を考慮したパラメータ生成

従来法 [1]では，DNN を学習する際にネットワークから出力された静的・動的特徴量の二乗誤差を最小化するようにネットワークの重みを更新する．本稿ではこの手法を Minimum Mean Squared Error（MMSE）学習と表記する．しかし実際に最小化すべき誤差は，動的特徴量を考慮し MLPG によって生成された静的特徴量であるため，MMSE 学習では損失関数の設定が適切でないという問題点が存在する．そこで本研究では最終的な生成誤差に対する誤差を損失関数とする Minimum Generation Error（MGE）学習を発話動画生成に導入しその効果について検討した．MMSE 学習でネットワークの重みを更新する手法を **MMSE**，MGE 学習でネットワークの重みを更新する手法を **MGE** とそれぞれ表記する．

これら 2 手法によって生成された動画に対して，自然性に関する主観評価実験を行った．実験において被験者は，2 手法の動画を比較し自然性の高い方の動画を選択する試行を 10 回行った．被験者は計 8 名だった．図 1 にその結果を示す．横軸は各手法の選択率，エラーバーは 95%信頼区間を表す．この実験結果から，MGE 学習を導入することで，より自然性の高い

動画を生成できることが示された.

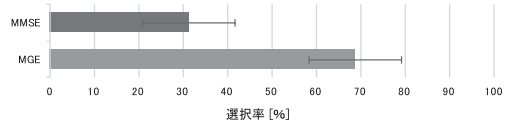

図 1 MMSE と MGE に対する主観評価結果

4. 系列内変動を考慮したパラメータ生成

DNN によって生成されたパラメータが過剰に平滑化し, 1 発話区間内で時間方向にパラメータを見たときの分散（系列内変動, Global Variance（GV））が小さくなってしまう過剰平滑化問題が, 音声合成に関する研究において問題とされている. 図 2 に **MGE** で生成したパラメータの GV の平均値をプロットした図を示す.

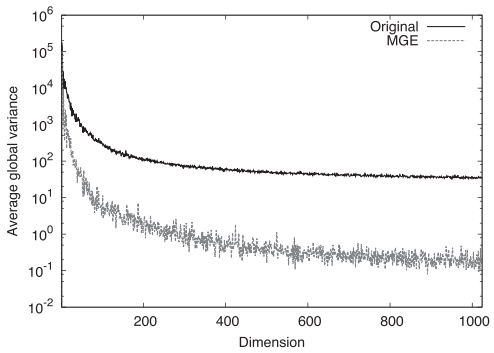

図 2 平均 GV の例

横軸は次元番号, 縦軸は平均 GV の値をそれぞれ表している. またプロットしたのは自然動画及び生成動画の AAM パラメータのうちのアピアランスパラメータであり, 黒い線が自然動画の, 赤い線が前章において **MGE** によって生成された動画の平均 GV を表す. 図からわかる通り, 生成されたパラメータの平均 GV は正解と比べ大きく減少していることがわかる. このように発話動画生成においても過剰平滑化が生じていることがわかる. しかしこれまでの先行研究において, 過剰平滑化問題が発話動画生成に与える影響については検討されてこなかった.

過剰平滑化問題を改善する手法は様々提案されているが, 本稿では Generative Adversarial Networks (GAN) を用いる手法を導入する. GAN は識別モデルと生成モデルの 2 つのネットワーク構造を持ち, それぞれのモデルを敵対させながら学習することで生成器の性能向上を図る. GAN を用いた音声合成に関する先行研究において, 従来の MGE 学習を用いた場合と比較して GAN の構造を導入することで生成パラメータの過剰平滑化が改善することが示されている [2]. 従って本研究では, GAN の構造を 3 章で検討した MGE 学習に導入しその効果を検討した. 簡略化のため, MGE 学習に GAN の構造を導入しパラメータを生成する手法を **MGE+GAN** と表記する.

MGE と **MGE+GAN** の 2 手法に対して, 客観・主観評価を行った. 客観評価では, それぞれの手法によって生成されたパラメータの平均 GV を比較した. その結果を図 3 に示す. 横軸は次元番号, 縦軸は平均 GV の値をそれぞれ表している. またプロットしたのは自然動画及び生成動画の AAM パラメータのうちのアピアランスパラメータである. この図より, **MGE** と比較して **MGE+GAN** ではパラメータの分散が大きくなり過剰平滑化が改善していることがわかった.

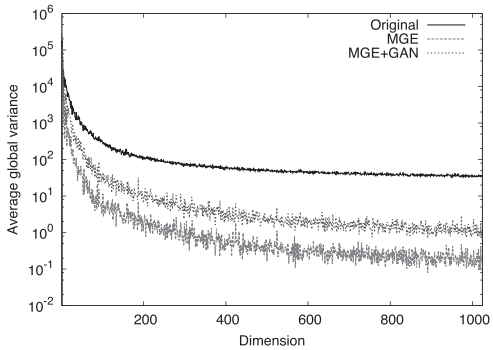

図 3 各手法における平均 GV の比較

次に, これら 2 手法によって生成された動画に対して, 自然性に関する主観評価実験を行った. 実験手法は 3 章での主観評価と同様である. 図 4 にその結果を示す. 横軸は手法の選択率, エラーバーは 95%信頼区間を表す. この実験結果より, GAN の構造を導入することでより自然性の高い動画が生成できることがわかった.

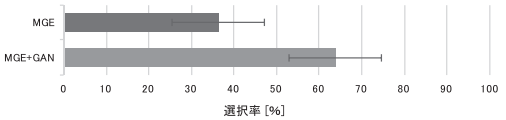

図 4 MGE と MGE+GAN に対する主観評価結果

5. まとめ

本研究では自然な発話動画生成を目指し, 従来手法では検討がされていなかった MGE 学習や GAN を用いたパラメータ生成手法について検討を行った. 主観評価を通じ, MGE 学習に GAN の構造を導入することで従来手法よりも自然な動画を生成できることを示した.

文献

1) J.Parker, R.Maia, Y.Stylianou, and R.Cipola, “Expressive visual text to speech and expression adaptation using deep neural networks,” ICASSP, 2017.
2) Y.Saito, S.Takamichi, and H.Saruwatari, “Statistical parametric speech synthesis incorporating generative adversarial networks,” IEEE/ACM TASLP, 2017.